ACOUSTIC ANALYSIS OF THE STOP CONSONANTS FOR DETECTING HYPERNASAL SPEECH

G. Castellanos, F. A. Sepúlveda *

Universidad Nacional de Colombia
Dpto. de Ingenierías Eléctrica,
Electrónica y Computación
Km 6 vía al Magdalena, Manizales
e-mail:fasepulvedas@unal.edu.co

Juan I. Godino-Llorente

Universidad Politécnica de Madrid
Dpto. de Ing. de Circuitos y Sistemas
Ctra. de Valencia Km. 7, 28031,
Madrid, Spain

ABSTRACT

Speakers having evidence of a defective velopharyngeal mechanism produce speech with inappropriate nasal resonance (hypernasal speech). Voice analysis methods for the detection of hypernasality commonly use vowels and nasalized vowels. However, to obtain a more general assessment of this abnormality it is necessary to analyze stops and fricatives. This study describes a method for hipernasality detection analyzing the unvoiced Spanish stop consonants */k/* and */p/*, as well. The importance of phoneme-by-phoneme analysis is shown, in contrast with whole word parametrization which may include irrelevant segments from the classification point of view. Parameters that correlate the imprints of Velopharyngeal Incompetence (VPI) over voiceless stop consonants were used in the feature estimation stage. Classification was carried out using a Support Vector Machine (SVM), obtaining a performance of 74% for a repeated cross-validation strategy evaluation.

1. INTRODUCTION

The speech communication process requires a translation of thoughts into spoken language. A person with a physical and/or neurological impairment may have a compromised vocal tract configuration and/or excitation, resulting in reduced speech quality. An specific example of a vocal tract dysfunction that reduces the speech quality, is the defective of velopharyngeal mechanism [1], which can be caused by one or more of the following factors: 1) physical defects (cleft palate), 2) central nervous system damage (traumatic brain injury), 3) peripheral nervous system damage (Moebius syndrome), and 4) impaired hearing [2]. The term cleft palate refers to a malformation affecting the soft and/or hard palate, and is usually congenital. This may result in reduced quality and clarity of speech.

The speech signal is mainly affected in two directions: 1) nasalized phonemes, due to a moderate to large velopharyngeal opening; and 2) weak consonants and short utterance length, due to loss of air pressure with air nasal air emission [3]. The most common approach to detect velopharyngeal disfunction (employing Digital Voice Processing, DVP) is by carrying out an analysis of vowels and nasalized sounds. In [2], a group of delay-based signal processing technique is described for the analysis and detection of hypernasal speech. Experiments were carried out with the nasalized vowels of the data */summer/*, */sunny/*, and */singing/* uttered by 33 hypernasal speakers and 30 normal speakers. Using the group delay-based acoustic measure, the performance on a hypernasality detection task is found to be 100% for */a/*, 88.78% for */i/* and 86.66% for */u/*. Furthermore, the effectiveness of this acoustic measure was cross-verified on data collected in an entirely different recording environment. In [1], the sensitivity of the Teager energy operator for multicomponent signals was used to detect hypernasality. A measurable difference was observed between the low-pass and band-pass profiles of the Teager energy operator for the nasalized vowels, whereas the normal vowel, which is a single component signal, does not show any difference. Additionally, in [4], a method based on the acoustic measures on several voiced phonemes was presented; however, a priori information about the velopharyngeal phenomenon was not used. Nonetheless, in the treatment of problems related to VPI several kinds of phonemes are used by the specialists. Not only vowels and nasalized vowels are tested, as considered in above mentioned works, but also the acoustic behavior of stops and fricatives are examined in order to assess the resonance problem, as well. Hypernasality detection by means of unvoiced stop consonants analysis using DVP requires the use of parameters that represents its acoustic behavior. Acoustic parameters for the analysis of pathological voices such as Harmonics to Noise Ratio (HNR), Normalized Noise Energy (NNE), Glottal to Noise Excitation (GNE), and so on, have been recently developed; however they were mainly designed to work for sustained vowels, provoking that they can not be used in this study. Parameters that correlate the imprints of Velopharyngeal Incompetence (VPI) over voiceless stop consonants were looked for and used in the feature estimation

*This work was supported in part by the scolarship 1989-2006 warded to F. A. Sepúlveda by COLCIENCIAS.

stage. Classification is carried out using a Support Vector Machine (SVM), and for classifier evaluation a repeated cross-validation procedure is used.

2. MATERIALS AND METHODS

First of all, it is necessary to take into account the drawbacks provoked by small training samples in the design of automatic classification systems. To reduce these problems, features used must correlate the influence of velopharyngeal incompetence in stop consonants, and classifiers with good generalization properties should be employed [5].

2.1. Database

The sample was made up of 88 children. Classes are balanced (44 patients with normal voice and 44 with hypernasality), and all registers were evaluated by specialists. Each recording contained several Spanish words, but in this study only the words "coco "*(/kóko/)* and "papá "*(/papá/)* were used. Signals were acquired under low noise conditions using a dynamic, unidirectional microphone (cardioid). The dynamic range of the signals was normalized between (-1, 1). A manual segmentation process was carried out to separate the stop phonemes (*/k/* and */p/*) of the utterances */kóko/* and */papá/* resulting in various sets, each formed by 88 signals.

2.2. Parametrization of plosive signals

A plosive consonant is formed by blocking the oral cavity at some point. During the articulation of most plosives the velum is raised, blocking off the nasal passages. This allows a certain amount of pressure to build up in the oral cavity behind the occlusion; if it were not raised, any pressure the speaker attempted to create behind the occlusion would leak through the nasal passage [6]. Plosives are produced by building-up and sudden release of oral pressure, requiring closure of the nasal passages with the velum. Individuals with cleft palate have never learned to control the movements of the velum, since, even with the velum raised, air pressure escapes trough the cleft into the nasal cavity. After reconstructive surgery or the fitting of a prothesis, such individuals need guidance in controlling the velum to produce plosive sounds [6].

The subglottal pressure represents the energy immediately available for creating the acoustic signals of speech. Innapropiate levels of subglotal pressure or innadecuate pressure regulation can cause abnormal speech intensity levels [7]. The pressure that is builds up behind the occlusion is released suddenly as a minor *explosion* or *popping* [6] [3]; thus, the measured *power* of stops may help to perceive the weakness of plosive consonants in velopharyngeal patients. In this study it is calculated using the expression:

$$P = 10 \log_{10} \frac{P_{\text{stop}}}{P_{\text{word}}} \quad (1)$$

where $P = \frac{1}{T_s} \sum_i x_i^2$ is the calculated power of the signal x, which can be the stop segment or the word, T_s is the size of support of x and i is the discrete time index.

Air leakage around the blockage will significantly slow down the supraglottal pressure, and therefore, the phonatory delay will drop down; the greater the shunt airflow, the larger the delay should be [7]. This can provoke a short utterance length of consonant plosives, which in this study, the duration is measured in relation to the word by means of the expression $D = T_{\text{stop}}/T_{\text{word}}$.

Velum action allows the nasal cavities to be closed or opened (or partially open, or perhaps air leakage around the velum blockage) with respect to the rest of the vocal tract, which allows sound waves to resonate within the nasal cavities, giving a distinctive nasal quality to the speech sounds produced [6]. In addition, the lower pressure of voiced stops in hypernasal speech results in less high frequency energy during the burst [7]. The MFCC and DWT (Discrete Wavelet Transform) use filterbanks to obtain measures of different portions of the spectrum, so the energies at every filter could be used to model the behavior at different ranges of frequency.

MFCC's are currently one of the most widely used features for Automatic Speech Recognition (ASR). in this study, these features are calculated for each unvoiced stop phoneme taking the discrete cosine transform of the logarithm of energy at the output of a Mel filter. In feature extraction processes based on the Fourier transform, the features that are extracted have fixed time frequency resolution because of the inherent limitation of the FFT. For this reason, classification of some phonemes, specifically stops, is difficult using these features. More recently, discrete wavelet transform (DWT) and wavelet packets (WP) have been used for feature extraction, because of their multiresolution capabilities [8]. In this study, a 6-level decomposition wavelet transform was tested to go over the mentioned problem.

Assuming the phonemes are nearly stationary, features extraction for the whole stop segment is applied such that two groups of features were obtained for each word analyzed. The first group is formed by the power, the duration and the 13 MFCC coefficients, and the second is constituted by the power, the duration and the 7 energy values (6 for detail band and 1 for the approximation band) of the 6-level decomposition of the wavelet transform using the 3_{th} order *daubechies* mother wavelet.

Feature sets are organized as follows, the set ξ_a^b consists of the power, the duration and the 13 MFCCs coefficients, where a can take the values of c and p depending the word analyzed ("coco "or "papá ") and b can be 1_{st}, 2_{nd} according to the analyzed stop segment of the word, the first or the second. The collection of the the power, the duration and the 7 energy values of the wavelet bands ζ_a^b is similarly organized as ξ.

2.3. Support Vector Classifiers

Support Vector Machines (SVMs) are used in this study principally for two reasons: SVMs have a relatively good

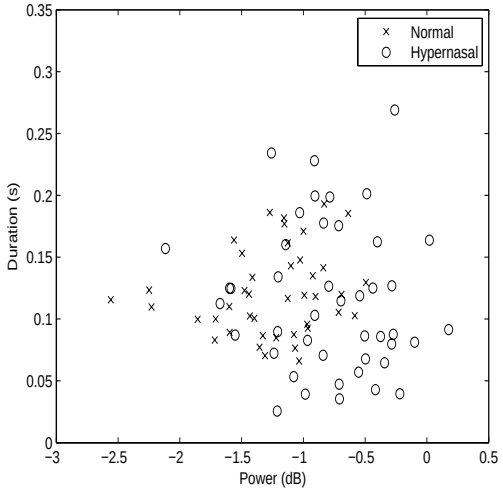

Figure 1. Duration vs power for the first plosive segment in the Spanish word */kóko/*

generalization capability with less amount of training data, and they have been particularly well developed for binary classification tasks. Traditional neural network approaches are more likely to suffer of poor generalization, producing models that can overfit the data [9], this is a consequence of the optimization algorithms used for parameter selection and the statistical measures used to select the "best" model.

For the binary classification problem we seek a discrimination function of the form [10]

$$g(x) = w^T \phi(x) + w_0 \tag{2}$$

with decision rule

$$w^T \phi(x) + w_0 \geq 0 \ \rightarrow x \in \omega_1 \tag{3}$$
$$w^T \phi(x) + w_0 \leq 0 \ \rightarrow x \in \omega_2 \tag{4}$$

where $\phi(x) : \Re^{n_1} \longmapsto \Re^{n_2}$ is generally a nonlinear function which maps vector x into what is called a feature space of higher dimensionality (possibly infinite) where classes are linearly separable. The vector w defines the separating hyper-plane in such a space and w_0 represents a possible bias.

The reason that makes SVMs more effective than other methods based on linear discriminants is its learning criterion. The goal of any classifier must be to minimize the number of misclassifications in any possible given sample. This is known as Risk Minimization (RM). However, in typical classification problems we only have a limited number of available samples (in some cases we can have an unlimited number of samples but, in any case, we only can deal with a subset), and thus, all we can do is to try to minimize the number of misclassifications within the training set. This is known as Empirical Risk Minimization (ERM), and most classifiers base their learning procedure on it [9].

However, having the classifier with the best ERM is not enough (or even desirable). The complexity of the classifiers must normally be fixed a priori, and so, we can end up having a too simple structure incapable of modelling correctly the classification boundaries of our problem, or a too complex one, overfitted to our training set and unable to generalize unseen example. This is known as Structural Risk, and a good classifier must maintain a compromise between the ERM and the SRM. In SVMs, one of the best advantages is that we do not need to fix the complexity of the resultant machine a priori. What we need is to fix a parameter which establishes this compromise between ERM and SRM [9].

3. RESULTS AND DISCUSSION

The utterance */kóko/* has two plosive segments, in the figures (1 and 2), 2-dimensional scatter plots using the duration and power for each segment can be observed. Discrimination

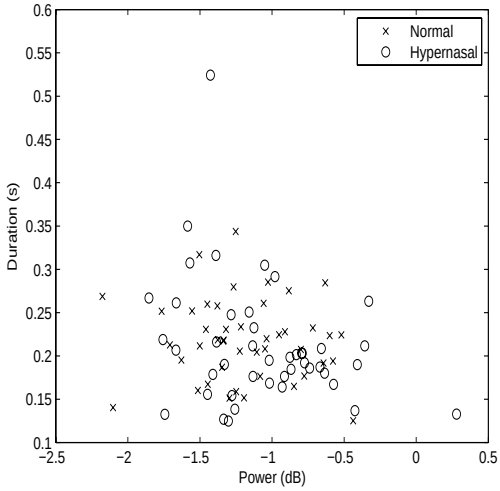

Figure 2. Duration vs power for the second plosive segment in the Spanish word */kóko/*

between the two classes can be observed using the first plosive segment ξ_c^{1st}, by contrast, on the second figure (obtained using the first and second features of the set ξ_c^{2nd}) this configuration can not be seen. The closure of the velopharyngeal gap is necessary to produce vowels as well as stops well; but in the first segment the velopharyngeral gap begins open, provoking in hypernasal children a delay in the closing phase. When beginning the second stop production, the velum is closed since the previous phoneme is a vowel, thus the behavior in relation to the duration is more similar to the normal category, as depicted by figure (2). A similar behavior is observed for the spectral features. Separability between the classes can be seen in the energy-bands scatter plots, nevertheless when this parameters are evaluated (joined to the *Power* and *Duration* inside the feature sets ξ_c^{1st} and ξ_p^{1st}) from the point of view of the classification rate, the best classifier's performance only

Table 1. Classification (%) results for the words */coco/* and */papá/* using the feature sets ζ_c^{1st}, ζ_p^{1st} and two kinds of kernels

Kernel	Quadratic		Radial Basis	
word	*/koko/*	*/papá/*	*/koko/*	*/papá/*
mean	63.52	unbounded	59.22	48.22
variance	5.12	unbounded	4.87	3.84

Table 2. Classification (%) results for the words */coco/* and */papá/* using the feature sets ξ_c^{1st}, ξ_p^{1st} and two kinds of kernels

Kernel	Quadratic		Radial Basis	
word	*/koko/*	*/papá/*	*/koko/*	*/papá/*
mean	72.78	73.85	63.89	56.15
variance	3.35	4.36	2.10	1.84

reached 63%. However, when 13_{th} order MFCC coefficients were used, instead of DWT, the performance goes up to 74% evaluated by applying a cross-validation strategy for 30 runs.

4. CONCLUSIONS AND FUTURE WORK

From the experiments can be concluded that hypernasal assessment should be determined analyzing phoneme by phoneme, instead of complete words. The acoustic properties of the same phoneme can be completely different in different parts of the uttered word due to variability on the behavior of articulators which depend so much from the context.

A preliminary set of experiments were described in which the features are based on the behavior of velopharyngeal mechanism in cleft palate kids. A performance of 74% was obtained over the voiceless plosive $/p/$ using the MFCCs, power and duration in the feature estimation stage; however, to formulate parameters that correlate in a better way the acoustic imprints of VPI over the several kinds of phonemes is important.

Due to in the treatment of problems related to VPI various kinds of consonants are used by the specialists, besides vowels, the fusion of hypernasality analyzing techniques on consonants, such as stops and fricatives, and vowels may give a better accuracy and confidence of the results. In addition, given that the phoneme by phoneme analysis is important and the manual segmentation is a time-consuming process, automatic segmentation should be considered inside the whole system for automatic hypernasal assessment.

5. ACKNOWLEDGEMENTS

This work was carried out under grants: 20201004208 funded by Universidad Nacional de Colombia; "Automatic detection of hypernasality in children with cleft lip and palate by means of acoustic analysis of speech" financed by DIMA; and TEC2006-12887-C02 from the Ministry of Science and Technology of Spain.

6. REFERENCES

[1] D. Cairns, J. Hansen, and J. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear teager energy operator," in *ICSLP-96:Inter. Conf. on Spoken Language Processing*, vol. 2, 1996, pp. 780–783.

[2] P. Vijayalakshmi, M. Ramasubba, and D. O'Shaughnessy, "Acoustic analysis and detection of hypernasality using a group delay function," *IEEE Trans. on Biomedical Engineering*, vol. 54, no. 4, pp. 621–629, April 2007.

[3] A. Kummer, *Cleft Palate and Craniofacial Anomalies: Effects on Speech and Resonance*. Clifton Park, NY: Thomson Delmar Learning, 2001.

[4] G. Daza, L. Sánchez, A. Sepúlveda, and G. Castellanos, "Acoustic feature analysis for hypernasality detection in children," *To be Published by IDEA Group Inc. in Encyclopaedia of Healthcare Information Systems*, 2008.

[5] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. on Pattern Analysis and Machine In telligence*, vol. 22, no. 1, pp. 4–36, January 2000.

[6] I. R. MacKay, *Phonetics: the science of speech production*. Allyn and Bacon, 1987.

[7] R. J. Baken, *Clinical Measurement of Speech and Voice*. Singular Publishing Group, Inc., 1996.

[8] O. Farooq and S. Datta, "Phoneme recognition using wavelet based features," *Information Sciences*, vol. 150, pp. 5–15, 2003.

[9] R. Solera-Ureña, D. Martín-Iglesias, A. Gallardo-Antolín, C. Pelaéz-Moreno, and F. D. de María, "Robust asr using support vector machines," *Speech Comunication*, vol. 49, pp. 253–267, 2007.

[10] A. R. Webb, *Statistical Pattern Recognition*. John Wiley and Sons, Ltda, 2002.